HHS Public Access

Author manuscript

Int J Audiol. Author manuscript; available in PMC 2015 December 03.

Published in final edited form as:

Int J Audiol. 2014 March ; 53(0 2): S5–15. doi:10.3109/14992027.2013.857435.

SHORT-TERM VARIABILITY OF PURE-TONE THRESHOLDS OBTAINED WITH TDH-39P EARPHONES

Gregory A. Flamme[a], **Mark R. Stephenson**[b], **Kristy K. Deiters**[a], **Amanda Hessenauer**[a], **Devon VanGessel**[a], **Kyle Geda**[a], **Krista Wyllys**[a], and **Kara McGregor**[a]

[a]Department of Speech Pathology and Audiology, Western Michigan University, Kalamazoo, MI, USA

[b]Division of Applied Research and Technology, National Institute for Occupational Safety and Health, Cincinnati, Ohio, USA

Abstract

Objective—To estimate the short-term variability and correlates of variability in pure-tone thresholds obtained using audiometric equipment designed for occupational use, and to examine the justification for excluding 8 kHz as a mandatory threshold in occupational hearing conservation programs.

Method—Pure-tone thresholds and other hearing-related tests (e.g., noise dosimetry, otoscopy, middle ear assessment) were conducted with a group of 527 adults between 20 and 69 years of age. A total of five measurement visits were completed by participants within a 14-day period.

Results—The 50 % critical difference boundaries were −5 and 0 dB at 4 kHz and below and −5 and 5 dB at 6 and 8 kHz. The likelihood of spurious notches due to test-retest variability was substantially lower than the likelihood of failing to detect a notched configuration when present. Correlates of variability included stimulus frequency, baseline threshold, acoustic reflectance of the ear, average noise exposure during the previous 8 hours, age, and the tester's level of education in audiology.

Conclusion—The short-term variability in 8 kHz pure tone thresholds obtained with the TDH-39P earphone was slightly greater than at other frequencies, but this difference was not large enough to justify the disadvantages stemming from the inability to detect a 6 kHz notch.

Keywords

Audiometry; noise-induced hearing loss; reliability; occupational health

Correspondence to: Gregory A. Flamme, Department of Speech Pathology and Audiology, Western Michigan University, 1903 W. Michigan Ave., Kalamazoo MI, 49008, greg.flamme@wmich.edu.

Preliminary findings from this study were presented at the 2013 National Hearing Conservation Annual Conference in St. Petersburg, FL.

Disclaimer

The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

Audiologists and physicians supervising the audiometric monitoring component of hearing conservation programs often must make judgments about whether a hearing impairment is *more likely than not* related to a person's work. Although insufficient on its own, audiometric configuration plays a large role in these judgments (Coles et al., 2000).

Notches, or "dips," in the audiogram have been observed since the beginning of the era of modern testing (Guild, 1932). Although no single accepted definition exists (Rabinowitz et al., 2006), notched audiometric configurations are characterized by poorer hearing sensitivity in the region of 3 to 6 kHz than surrounding frequencies. Improved hearing sensitivity on the high-frequency side of the notch is a necessary component of notch identification.

Notched audiometric configurations are common among people who have had excessive noise exposure (Gravendeel & Plomp, 1959); although other conditions can also produce notched configurations, a notched configuration is often interpreted as an indication of noise-induced permanent threshold shift among people known to have high levels of occupational noise exposure (Coles, et al., 2000).

Bulged audiometric configurations (Dobie, 2005) are also indicative of excessive noise exposure. A bulged configuration is characterized by thresholds between 1 kHz and either 6 or 8 kHz that are substantially worse than a line drawn between the frequency limits. Bulged configurations can be identified in the absence of improvements in pure tone thresholds at the highest frequency tested. It is possible that audiometric configurations that present initially as notches transform to bulges over time as improved sensitivity on the high-frequency side of the notch is minimized by age-related deterioration (ANSI, 1996).

The effects of aging on the auditory system (ANSI, 1996) lead to monotonically decreasing hearing sensitivity with increased frequency. Thus, the presence or absence of improved sensitivity on the high-frequency side is an important source of information for ruling out noise as a potential contributor to an observed threshold shift.

It is commonly assumed that audiometric notches occur more commonly at 4 kHz than at any other frequency, based on noise-induced threshold shift data (ANSI, 1996). However, evaluation of NHANES data (Stephenson, et al. 2004; Stephenson & Flamme, unpublished analyses of NHANES 1999–2004 data) revealed that 6 kHz was the most affected frequency in approximately 50 % of audiometric notches found among people between the ages of 20 and 69 who report having been exposed to occupational noise for 3 months or more (Figure 1). In order to minimize bias from potential calibration errors (Lutman & Qasem, 1998), the notch definition used in these analyses was conservative, requiring a 15 dB notch depth relative to the mean threshold at 0.5, 1, and 2 kHz, and a 15 dB recovery or improvement at the 8 kHz stimulus frequency from the worst threshold frequency. The prevalence of notches centered on 6 kHz is at least as great as the prevalence at lower frequencies, so it is crucial that test protocols permit their detection.

The OSHA Hearing Conservation Amendment (OSHA, 1983) requires annual audiometric testing to support early identification and remediation of occupationally-induced threshold shifts, but neglects the recommendation to include 8 kHz due to three concerns: 1)

calibration stability at 8 kHz, 2) “spurious results” (i.e., reliability) at this frequency and 3) the fact that some audiometers in use at the time were incapable of testing at 8 kHz (OSHA, 1981). The concern about audiometers incapable of testing at 8 kHz is no longer relevant because current audiometers are designed to test up to 8 kHz.

In 2003, the Department of Labor implemented new requirements for recording occupational illness and injury. For the first time, occupational hearing loss became a recordable illness on OSHA’s Form 300 (OSHA 2003). However, a hearing loss must be work-related in order to be recordable. Because NIHL is often characterized by the presence of a noise notch, the absence of a notch makes it very difficult to judge a hearing loss as being noise-induced. Without data at 8 kHz, the noise-induced configuration could not be differentiated from presbyacusis, or much more rarely, retrocochlear pathology. The final outcome in this chain of events would be that the noise-induced hearing loss would not be considered “work-related” and therefore would not be recorded on an OSHA Log 300. Consequently, the affected individual would be unaware of the probable cause of his/her hearing loss, the worker’s employer would be unaware there was an occupational illness, and the federal or state health and safety regulators would be unaware of the problem.

Audiologists and physicians involved in the interpretation of audiometric configurations must remain cognizant of the inherent variability in threshold measures. The study of test-retest variability of pure tone thresholds has a long history (Carhart & Hayes, 1949; Atherley & Dingwall-Fordyce, 1963; Dobie, 1983), and some have argued that the high variability of pure tone thresholds obtained in occupational settings compromise the usefulness of such data (Hetu, 1979; Atherley & Johnston, 1981). This position is inarguable in cases where inconsistent procedures, inadequate test environments and hardware, etc. fail to produce interpretable threshold data, but the presence of flawed data when better data can be had represents a call for better technique rather than the dismissal of testing altogether.

Key studies on test-retest variability in occupational audiometry were done by High and Gallo (1963) and Dobie (1983). In the High & Gallo study, test-retest standard deviations between the frequencies of 0.5 and 6 kHz ranged between approximately 4 and 6.5 dB, with the greatest variability observed at 4.0 kHz. In the Dobie study, a database of over 4500 workers was reviewed with respect to the expected distribution of test-retest differences in pure tone thresholds and alternate definitions of Standard Threshold Shifts involving different frequencies, etc. Test-retest difference distributions were observed over a period of 6 – 18 months, and standard deviations of threshold differences at individual frequencies ranged between 6.8 dB (2 kHz) and 10.0 dB (6 kHz).

Hearing impairments associated with aging tend toward monotonically increasing thresholds with frequency, while impairments associated with exposure to noise and other factors tend to be notched or bulged, with the most affected in the region of 3 – 6 kHz. Analyses of data from the National Health and Nutrition Examination Survey (NHANES) revealed that in most cases, the worst threshold in a notched configuration is observed at 6 kHz (see Figure 1) (Stephenson et al., 2004, Flamme & Stephenson, unpublished analyses, 2013), which requires measurements at 8 kHz to inform interpretation of the impairment as related to exposure, aging, or their combination. Threshold tests at 8 kHz are not mandatory in the

OSHA Hearing Conservation Amendment (OSHA, 1981) and this limitation reduces the interpretability of poor thresholds at 6 kHz because notched and bulged configurations centered on 6 kHz cannot be identified.

The current study was undertaken to determine whether exclusion of the 8 kHz audiometric test frequency was justified by increased test-retest variability, and identify factors related to test-retest differences.

Methods

Participants

Participants were 527 (52 % male) non-institutionalized adults between the ages of 20 and 69 years at the time of enrollment. Participants were drawn from the general population in and around Kalamazoo, Michigan between the years of 2009 and 2012. The target sample size was 50 participants of each gender and decade of age, as determined by power analyses conducted *a priori* assuming separate analyses of each combination of gender and age.

To be included in the study, participants were required to have pure tone hearing thresholds better than 80 dB HTL at all stimulus frequencies from 0.5 to 8 kHz, and no more than 40 dB of asymmetry at any stimulus frequency on at least one test during the first measurement visit. We also required visibility of the tympanic membrane by conventional otoscopy without interference of cerumen, no signs and symptoms of middle ear dysfunction via otoscopy and conventional tympanometry, and sufficient capacity to understand written and spoken instructions and study materials. Participants were excluded from participation if they failed a hearing screening at 70 dB HTL during the enrollment visit or if they were unable to perform the modified Hughson-Westlake threshold measurement task via automatic, semi-automatic, or manual test procedures.

The participant flow diagram is represented in Figure 2. The vast majority of participants (85 % of men, 89 % of women) enrolled in the study completed the study protocol. Around 6 % of enrolled participants (16 men and 17 women) withdrew from the study, mainly due to unforeseen scheduling complications. About 9 % of men and 4 % of women were dismissed from the study after enrollment. In most cases (16 men, 4 women), participants were dismissed because they were unable to hear all tones during the hearing screening conducted during the enrollment visit, or because of cerumen that interfered with the investigator's view of the tympanic membrane. Other common reasons for dismissal were threshold asymmetry (7 men, 1 woman) detected during the first measurement visit and signs or symptoms of middle ear dysfunction that were detected during a measurement visit and not resolved during the 14-day time frame for measurement visits (one man, four women).

Participants were asked complete a background information form at the time of enrollment and another form detailing recent exposures and events at the beginning of every measurement visit. These forms included items concerning educational attainment, current (or most recent) occupation, history of occupational and non-occupational noise exposure and hearing protector use, hearing status, history of ear infections, pressure equalization

tubes/grommets, tinnitus, dizziness, and history of smoking, and the amount of time that had passed since the participant's most recent hearing test. Information obtained at the beginning of each measurement visit included recent changes in hearing, current tinnitus or dizziness, ear pain, exposure to sound through personal music players, and exposure to potentially ototoxic chemicals (e.g., cigarette smoke, paint, liquid adhesives, pesticides, fuels).

The educational attainment of participants included all educational categories from less than grade 11 in secondary school to earned Doctoral degrees. The predominant levels of educational attainment were earned Bachelor's degrees (26 %), some college (19 %), high school graduates (16 %), and awarded Master's degrees (14 %). The occupations of participants covered the range of major Standard Occupational Classifications developed by the US Bureau of Labor Statistics (US Department of Labor, 2010), with the exception of military service. The most common participant occupations were in the Office/ Administrative Support (17 %), Education/Training/Library (13 %), Production (7 %), and Building/Grounds Maintenance (7 %).

Human research subject protection oversight was provided via review boards at WMU and NIOSH (protocol HSRB 09-DART-05XP). Participants were provided reimbursement for their time and inconvenience at the end of each completed measurement visit. The total reimbursement for participants completing the study was $77, which corresponds to approximately $13 per hour spent on study-related activities.

The data collection team for this study met the OSHA criteria for personnel involved in occupational hearing testing, including audiologists, students currently enrolled in graduate-level audiology training, and undergraduate students with training in hearing science and hearing test procedures comparable to that for certified Occupational Hearing Conservationists (OHC). Audiologists and students enrolled in graduate-level audiology training conducted all procedures as needed, while the role of testers with OHC-level training was limited to audiometry, measurements of ear canal collapse, and daily audiometer calibration checks.

Instrumentation

Hearing screening was conducted using pure tone stimuli (2, 3, 4, 6, and 8 kHz, stored on a personal music player) delivered at 70 dB HTL (ANSI S3.6 - 2004) using Sennheiser® HDA200 (Sennheiser Electronic Corporation, Old Lyme, CT) circumaural earphones. Pure tone thresholds were assessed using the Tremetrics® HT Wizard audiometer (Tremetrics, Inc., Eden Prairie, MN), which was equipped with TDH-39P earphones with an HB-7 headband and MX41A/R cushions. The coupling force of the HB-7 headband was monitored and remained stable at approximately 4.7 Newtons throughout the study. The audiometer was designed for occupational audiometry and had the capacity for automated threshold testing using a modified Hughson-Westlake procedure (Carhart & Jerger, 1959). Pure tone testing was conducted in a double-walled sound booth with ambient noise levels permitting testing to -10 dB HTL (ANSI, 1999) at all stimulus frequencies.

Audiometric signals were calibrated using the GRAS Type 43AA ear simulator, which meets IEC 60318-1 (IEC, 2009) specifications. The SPL developed in the ear simulator was

measured using the Larson Davis System 824 sound level meter, which meets ANSI S1.4 Type 1 specifications (ANSI, 1983). The audiometer was calibrated according to ANSI S3.6-2004 (ANSI, 2004), which was current at the time data collection began.

Middle ear assessments consisted of conventional tympanometry (226 Hz) and wideband reflectance measures. Conventional 226 Hz tympanometry was conducted using the Interacoustics® MT10 (Interacoustics, Assens, Denmark) and Maico MI 24 (Maico Diagnostics, Eden Prairie, MN), which are tympanometers designed primarily for screening. Most data were collected using the Interacoustics MT10 because a custom MATLAB® routine was available for automatic transfer of tympanogram data to a computer. Wideband reflectance (WBR) represents the amount of energy reflected from the eardrum as a function of frequency. Its complement is wideband energy absorbance. Wideband reflectance at ambient pressure and wideband tympanometry (WBT) were assessed using the Interacoustics WBR/WBT (research-oriented) system. This system relies on a Windows-based PC system, which drives a sound card and a specially-modified Interacoustics impedance audiometer outfitted with an Interacoustics Titan probe.

Noise exposure was monitored using Etymotic Research® ER200D personal noise dosimeters. These devices were configured for an 85 dBL_{Aeq8} criterion, 3 dB exchange rate, 65 dBA threshold, and the devices are designed to accept inputs up to 130 dBA. Further details and noise exposure data from the first 286 participants in this study are presented elsewhere (Flamme et al., 2012).

Ear canal size and degree of collapse under earphones were made. Canal size measurements were made using the 3M™ Eargage ear canal sizing tool, which is a set of five ball gauges sized between 7.6 and 11.5 mm that were inserted into the ear canal with the objective of identifying the largest ball that fit comfortably into the ear canal entrance.

Ear canal collapse measurements made using disabled TDH-39P earphones with MX41A/R cushions mounted in a Telephonics HB-9 headband applying the same coupling force as the audiometric earphones. The electronics and back sections were removed from the earphone cases to permit visualization of the ear canal with an otoscope.

Procedure

Requests for volunteers for this research project were distributed via electronic and paper notices. Potential participants contacted the research team via email and telephone, and a brief summary of the study goals, procedures, and participant criteria was provided. Potential participants were then invited to set a meeting with an investigator. This meeting included a description of study details, verification of participant interest, documentation of informed consent, otoscopy to rule out excessive cerumen, pure tone hearing screening, administration of the enrollment form, issuance of the personal noise dosimeter, and scheduling of measurement visits.

Participants attended measurement visits at least 16 hours after enrollment in order to ensure 16 hours of continuous noise dosimetry prior to measurement of pure tone thresholds. The study protocol allowed for five measurement visits that all occurred within a 14-day time

frame. Each measurement visit included the same procedures, with the exception of the first ear examined, which was determined randomly prior to the first measurement visit. Testers were unable to access observations from any prior visits when making observations for the current visit.

Measurement visits took place in a research laboratory environment and followed a strict protocol typically lasting between 30 and 40 minutes. Measurement visits were separated by at least 16 hours and at most 7 days, and the final measurement visit could not occur more than 14 calendar days after the first measurement visit.

Upon arrival in the laboratory, participants were seated in the main laboratory room, returned their noise dosimeter for download by testers, and completed the daily information form. Otoscopy, measurements of ear canal size and middle ear assessment followed. The middle ear assessment procedure began with bilateral conventional tympanometric measures, followed by WBR measurements at ambient pressure and WBT measures. Six repetitions of WBR measurements and two repetitions of WBT measurements were completed with a given probe placement, with the tester repositioning the probe as necessary to obtain a pneumatic seal. An inadequate seal for the WBR test was indicated by low reflectance (e.g., < 0.65) below 250 Hz. An inadequate seal for the WBT test was indicated by failure to achieve the desired positive static pressure during the test. Although the reflected energy near ambient pressure was the primary objective of these tests, the protocol allowed the tester to reposition the probe and repeat the test a few times to increase the probability of a complete WBT record.

Participants were allowed momentary break and then took a seat in the sound booth. Ear canal collapse was then assessed using the disabled earphones described above. The disabled earphones were placed on the participant's ears with the tester standing in front and slightly to the side of the participant. The side on which the tester stood was selected based on the first ear tested. The tester lifted and replaced one earphone from the head while observing the change in the shape of the ear canal and judged the percentage of ear canal collapse at the widest point on the ear canal with the earphone removed. Judgments were limited to five equidistant categories ranging between 10 and 90 % collapse.

Audiometric thresholds were obtained after measurements of ear canal collapse were made. Instructions were read aloud to each participant while directing the participant's attention to a printed copy of the instructions. The same instructions, which were pre-recorded and stored on the audiometer, were then played via the earphones. Testing was then conducted for the selected ear via the instrument's (modified Hughson-Westlake) protocol, with semi-automatic (i.e., automated testing of a frequency and manual switching between frequencies), and manual testing conducted as needed.

Calibration checks of all instrumentation involving acoustic stimuli were conducted twice per day. Tympanometer calibration was checked using tympanometer test cavities. The manufacturer's calibration apparatus was used for WBR and WBT calibration. Audiometer calibration checks were conducted using the IEC-60318-1 ear simulator (IEC, 2009).

Data analyses

Noise dosimetry data were downloaded to spreadsheet format from the devices using the manufacturer's software. These data were then transformed into equivalent continuous A-weighted levels in the 1, 2, 4, and 8 hours prior to the participant's arrival for the measurement visit and the overall average sound level and cumulative noise dose during the period since dosimeter issuance or the previous measurement visit. These transformations were implemented using MATLAB (Mathworks, Inc., Natick, MA) software.

Conventional 226 Hz tympanograms and summary data (e.g., static compliance, tympanic peak pressure) were downloaded from the MT10.

Reflectance proportions from wideband absorbance measurements (ambient pressure and wideband tympanograms) were extracted from binary file format using manufacturer-supplied MATLAB functions. These data were then averaged across runs (i.e., six runs at ambient pressure, two wideband tympanograms) at each frequency. The mean reflectance was then transformed into dB re 1.0 (i.e., dB relative to 100 % energy reflectance) to match the dB scale used in audiometric measurements.

Audiometric data were downloaded to spreadsheet format from the HT Wizard Ultra audiometer via a serial interface. A custom MATLAB function was then used to collate threshold data across tests and display thresholds and threshold differences for each participant. Threshold differences were calculated using the first test on the first measurement visit as the reference. All other study data (information forms, otoscopy, procedure checklist variables) were entered into a Microsoft Access® (Microsoft, Redmond, WA) database prior to merging all data into a single dataset.

Descriptive analyses were conducted on all available data. Inferential analyses involved preliminary assessments of the degree and nature of correlations between test-retest differences obtained across stimulus frequencies followed by the identification of significant correlates of test-retest differences. The correlations between test-retest differences are important because significant positive correlations across a group of frequencies would suggest that thresholds at those frequencies tend to increase or decrease as a group, thus preserving the overall audiogram shape. Knowledge of the correlates of test-retest differences could be used to inform test protocols or the interpretation of test results. It was of interest to this study to identify the factors associated with apparent improvements and decrements in hearing sensitivity across repeated tests, and this implied a dependent variable that retains the sign of the test-retest difference.

Assessment of the correlation structure across stimulus frequency was conducted using Pearson (zero-order) correlation coefficients and structural equation modeling (SEM), both of which using functions implemented in Stata® (StataCorp, College Station, TX) software. The SEM fitting was conducted with standardized threshold differences (i.e., *z*-scores) to normalize variance across stimulus frequencies. Maximum Likelihood estimation was used to estimate parameter values. A jackknife procedure, clustered on the participant identifier, was used in the calculation of the standard errors for SEM model coefficients. Briefly, the jackknife procedure estimates variance of model coefficients based on the distribution of

changes in coefficients when each member of a cluster (e.g., participant identifier) is temporarily removed from the sample.

This study required multilevel modeling to account for the nested structure of the data. For example, observations of test-retest differences for one ear would be expected to be more strongly related to one another than differences across ears, owing to the possibility of temporary effects (e.g., middle ear ventilation) that could affect one ear but not the other. Similarly, observations within one test in a given visit were expected to be more closely related to one another than observations across tests. Observations within one visit and within one participant were also expected to be related. Thus, a 5-level model (i.e., observations nested within ears, ears nested within tests, tests nested within visits, and visits nested within participants) was used to represent this correlation structure in inferential analyses. Robust standard error calculations (Huber, 1967) were used to minimize the effects of any violated statistical assumptions that went undetected.

Although univariable models of the correlates of test-retest difference included more observations, the final multivariable model was based on threshold differences from baseline (n = 43651), which were nested within ear status at the time of test (n = 6239), tests (n = 3133), measurement visits (n = 1756), and participants (n = 441). The primary reason for the reduction in the number of participants was missing data associated with WBR/WBT system hardware failures and limited availability of replacement equipment during the period of data collection.

Audiometric, demographic, responses to informational items, noise exposure, otoscopic, middle ear, and tester variables were included as potential predictor variables, which were treated as fixed factors in the multilevel models. The development of multivariable models was informed by Hosmer and Lemeshow (2000), which suggested an initial screen for potential univariable predictors ($p < 0.2$) prior to the development of a multivariable main effects model. Significant predictors in the main effects model were then tested for interactions en route to the definition of the final model. Incomplete data from participants who withdrew or were dismissed were retained in these analyses.

Results

Baseline pure tone thresholds for participations (i.e., the observed threshold during the first test obtained during the first measurement visit) revealed good hearing sensitivity for most participants through 2 kHz and poorer sensitivity at higher frequencies (Figure 3). Although most baseline thresholds were better than 40 dB HL, some thresholds were greater than 45 dB HL. Selected descriptive results and indications of which variable showed potential univariable relationships with test-retest differences are presented in the Appendix included in the supplemental online materials.

The most common test-retest threshold difference was 0 dB, regardless of stimulus frequency, and the preponderance of threshold differences (> 94 % at all stimulus frequencies) were found within two audiometric steps (i.e., 10 dB) of the baseline threshold. Threshold differences beyond this range were occasionally found as much as 55 dB from the

baseline threshold (Figure 4), but large differences were rarely duplicated across other measurement visits and/or tests within a visit.

Due to the small number of audiometric steps that encompass the vast majority of test-retest differences, critical difference calculations were based on percentiles of the test-retest difference distribution rather than standard deviations and a hypothesized normal distribution. Descriptive statistics and critical differences indicating the boundaries that must be exceeded for 50, 80, and 90 % certainty that an observed difference is not due to chance are represented in Table 1.

Test-retest differences were correlated across frequency. The Pearson correlations between test-retest differences at individual frequencies range between r = 0.33 and r = −0.01 (Table 2). The correlation pattern tends toward the strongest relationships occurring between neighboring stimulus frequencies. However, the only non-significant correlation in Table 2 is between the 4 and 6 kHz stimulus frequencies (r = −0.01; p = 0.39), while the correlation magnitude between 6 and 8 kHz was among the strongest observed in these data. This correlation pattern suggested that there might have been more than one factor associated with the pure tone threshold differences observed in this study, and this possibility was explored via SEM. The SEMs fitted to the standardized pure tone threshold difference data were intended to assess the likelihood that test-retest differences in pure tone thresholds were associated with more than one common factor. This difference was represented by a correlation between latent factors representing the low stimulus frequencies (<= 4 kHz) and the high stimulus frequencies (> 4 kHz), a Wald test of the hypothesis of a perfect correlation existed between the low- and high-frequency common factor, and the difference in the goodness of fit between one model that allowed this correlation to be freely estimated and another that fixed this correlation at 1.0 (i.e., merging them into a single factor), as estimated using the Likelihood Ratio test.

The two-factor SEM results are represented in Figure 5. The single-factor SEM structure was identical with the exception that the correlation between the two common factors was fixed at 1.0, thus yielding a single factor. The estimated correlation between the Low Hz and High Hz common factors was 0.67 (95 % CI: [0.50, 0.85]), the Wald test of the hypothesis of a perfect correlation between the factors was rejected (F (1,489) = 13.45; p = 0.0003), and the Likelihood Ratio test of the differences in fit between the models containing one versus two factors revealed a significant improvement in fit for the model including two factors (χ^2 (1) = 56.8; p < 0.00005). All statistical indicators suggest that the high stimulus frequencies (6 and 8 kHz) tend to covary semi-indepenently from the other stimulus frequencies.

Probabilities of spurious 6 kHz notches

The rate of spurious 6 kHz notches from chance variation in thresholds is complicated by the frequency dependence of test-retest differences (Figure 4 and Table 1) and the presence of correlated deviations across frequency (Figure 5 and Table 2). A spurious 6 kHz notch would result from apparent improvements in thresholds on both sides of 6 kHz along with an apparent decrement in threshold at 6 kHz. To test this, an indicator variable was generated to detect all tests in which (1) at least one deviation from the baseline pure tone thresholds at 2, 3, or 4 kHz was less than or equal to −5 dB, (2) the deviation from the baseline pure tone

threshold at 6 kHz was greater than or equal to 5 dB, and (3) the deviation from the baseline pure tone threshold at 8 kHz was less than or equal to −5 dB. These would be conditions resulting in an erroneous 6 kHz notch having a 10 dB depth. Only tests meeting all three criteria were identified as a spurious 6 kHz notch.

The rates of spurious notches were calculated for each audiogram and for two consecutive audiograms within a measurement visit. The proportions of tests producing spurious 10 dB notches at 6 kHz were low (see Figure 6 for results from the left ear), with 0.08 being the highest boundary for the 95 % CI for the proportion (measurement visit 4, test 2). The rates of two consecutive spurious notches within a measurement visit were lower, with a highest boundary for the 95 % confidence interval of 0.03. Results were similar for both ears, with less than a 0.01 difference in probabilities across ears for each measurement visit and test.

Correlates of test-retest differences

In addition to the threshold difference distributions (Figure 4) and critical differences (Table 1), it is helpful to also identify the factors associated with test-retest differences. Such information can inform both the conduct and the interpretation of test-retest differences. In this study, we examined the factors associated with improvements and declines in apparent hearing sensitivity on retest (i.e., a signed difference). Many variables showed potential relationships in the univariable context (Appendix), but most failed to retain significance in multivariable analyses (Table 3). Two-way interactions with stimulus frequency were tested but failed to provide a significant ($p > 0.05$) improvement in model fit via the likelihood ratio test.

The final model contained significant fixed effects (Wald χ^2 (15) = 326.26; $p < 0.00005$). The fixed factors in the final statistical model consisted of stimulus frequency, threshold at baseline, wideband reflectance (dB) at baseline, change from baseline in wideband reflectance (dB), average noise exposure during the 8 hours preceding the measurement visit, decade of age, and the audiology education status of the tester obtaining audiometric thresholds. The residual errors in the model fit were approximately normally distributed, with two to four percent more residuals observed in the area of 0 dB (+/− 1 dB) and slightly more frequent (< 0.1 %) extreme values than would be expected of a normal distribution.

The coefficients in the statistical model indicate that reductions (i.e., apparent improvements) in pure tone thresholds were observed at the lowest stimulus frequencies (e.g., approximately 2 to 2.5 dB at 0.5 and 1 kHz), relative to 8 kHz. These differences were less than one decibel at 4 kHz and above.

The audiometric threshold observed at baseline was inversely related to observations of test-retest differences. Poorer thresholds were more likely to be associated with apparent improvements upon retest, and although the magnitude of this coefficient was small (−0.133 per dB) the effect is among the strongest observed in this study. This association represents a clear trend toward an increased likelihood of apparent improvement on retest over the large range of thresholds characteristic of high-frequency hearing impairment.

Controlling for the other factors in the analytic model, older participants were more likely to have apparent worsening of thresholds upon retest. The magnitude of this difference was roughly 1 to 1.5 dB of threshold increase per decade of age. All age groups were significantly different from one another in *post hoc* contrasts with Bonferroni correction, with the exception of participants in their 30s and 40s, who were not significantly different.

Wideband reflectance (WBR) bore a significant relationship with test-retest pure tone threshold differences, both in the baseline measures and changes relative to baseline that were observed at the time of pure tone threshold testing. Observations of greater WBR, expressed in dB, were associated with an increased likelihood of apparent worsening of pure tone thresholds on retest. This relationship was present with both baseline WBR and changes from the baseline, suggesting that apparent decrements in hearing sensitivity were more likely among participants with less efficient middle ear systems.

A direct relationship was observed between test-retest differences and the participant's average noise exposure (dB L_{Aeq}) during the 8 hours prior to the measurement visit. The direction of the relationship is reasonable, suggesting an apparent worsening of threshold among those with greater amounts of average noise exposure.

Observations made by testers with any graduate-level education in audiology tended to increase the test-retest difference by about 0.3 dB. This result suggests that tests done by people with audiology-level education had a greater likelihood of slightly worse thresholds on retest. *Post hoc* exploration of these effects at the level of individual testers revealed that although the effect of audiology education differed slightly across testers, the difference was present across the group and the addition of individual testers as predictors in the model neither improved the overall fit of the model (Likelihood ratio test $p > .05$) nor eliminated the significance of the overall education effect.

The random effects of participant identifier, measurement visit, and ear were significant. The factor representing tests within measurement visits was not significant, suggesting that test results obtained within a given measurement visit can be considered independent samples of a listener's hearing status. However, the significant random effects associated with participant identifiers, measurement visits, and ears indicate that measurements across these factors are not independent. The average effect of measurement visit was small, with mean effects across measurement visits of −0.64 dB at visit 2 (95 % CI: [−8.5, 5.0]) to −0.56 dB at visit 4 (95 % CI: [−9.0, 5.1]). The wide range of confidence intervals suggests that there were individual differences in responses to each measurement visit that weren't included in other aspects of the statistical model.

Discussion

The results of this study indicated that short-term test-retest differences obtained in 5-dB steps using an automated audiometer and TDH-39P earphones will *more likely than not* remain between −5 and 0 dB re: baseline values through 4 kHz and between −5 and 5 dB re: baseline at both 6 and 8 kHz (Table 1). These differences are intercorrelated across frequency, and differences at 6 and 8 kHz are more strongly correlated with one another

than at frequencies below 6 kHz (Figure 5). Stimulus frequency, baseline threshold, WBR, noise exposure, age, and the education of the tester were all associated with test-retest differences.

The standard deviations of the difference distributions observed in this study (Table 1) were considerably smaller than those reported by Dobie (1983) and Carhart and Hayes (1949), yet greater than those reported in Atherley and Dingwall-Fordyce (1963). In the audiometry portion of the study, we used personnel, equipment, and procedures that can be realized in industrial environments, with the possible exception of the low ambient noise in the test booth. The minimum testable thresholds in the current study's test environment were sufficiently low to rule out masking effects, but this would not often be the case in industrial environments if supra-aural earphones are used. Fluctuating ambient noise levels would be expected to increase the variability seen at the low- and middle-frequencies.

The relationship between baseline thresholds and test-retest differences in the current study (Table 3) concurs with prior findings (Dobie, 1983) (p. 913) illustrating that test-retest differences are compressed by the dynamic range of the audiometer and/or test environment. It is possible that testers will judge the high frequencies to be more likely to improve on retest because of both the frequency dependence of test-retest differences and the increased probability of poorer high frequency hearing sensitivity.

Pros and cons of excluding 8 kHz

The results of this study confirm that thresholds obtained with the TDH-39P earphone have greater test-retest variability at 8 kHz than thresholds obtained at lower frequencies (Table 1), and although steps should be taken to reduce this variability, this finding alone does not necessarily justify the exclusion of 8 kHz from the mandatory test protocol.

In order to justify the exclusion of 8 kHz from routine testing, the consequences of inclusion of 8 kHz thresholds must be worse than the consequences of exclusion, and the results of this study indicate that the consequences of excluding 8 kHz are much worse. The OSHA requirement for recording only work-related significant threshold shifts (OSHA, 2003) mandates a work-relatedness determination, and a considerable portion of this determination hinges on the identification of a notch. The majority of notches are found at 6 kHz among people reporting a history of occupational noise exposure (Figure 1) and these require measurements at 8 kHz to identify a notch. Therefore, the evaluator (audiologist or physician) would be disadvantaged frequently when making work-relatedness determinations without access to 8 kHz thresholds.

On the other hand, a spurious 6 kHz notch of 10 dB or more is unlikely to occur in more than about 8 % of cases on individual retests or about 3 % of consecutive tests within a follow-up measurement visit, assuming that the *true* configuration for the listener is flat. Although the exclusion of 8 kHz thresholds from the mandatory protocol eliminates the possibility of spurious 6 kHz notches, this benefit is conveyed by also eliminating the possibility of detecting the most common notches. Using a *more likely than not* criterion (Coles, et al., 2000), there is no rational justification for accepting a 50 % error rate (Figure 1) en route to avoiding what is, at most, an 8 % error rate (Figure 6).

Although the random effects of participant identification, measurement visit, test within the visit, and ear were included in the analytic model to manage the correlation structure of the data, the implications of these findings bear discussion. We found no evidence for a between-test correlation in these analyses, which suggests that tests taking place within a measurement visit can be combined into an unbiased average that better represents the listener's central tendency than either test alone. However, the significant effect of measurement visit suggests that an average across visits could obscure systematic differences across visits, and these would result in a biased average. This finding does carry practical implications, particularly given that the variance component associated with measurement visit is comparable to the variance component associated with ear status at the time of testing. It is a conventional recommendation in occupational testing to repeat testing within a 30-day period to confirm a Standard Threshold Shift (OSHA, 1983), with the expectation that a Standard Threshold Shift that was not persistent was likely due to the unintended effect of excessive noise exposure or random measurement error. The variance component associated with measurement visit in an analysis including the array of fixed factors included in the current study suggests that differences across measurement visits are not entirely random or associated with excessive noise exposure. Further study of differences across measurement visits could inform the interpretation of follow-up tests by including time of day (Veneman et al., 2013) as well as other factors.

Correlates of test-retest differences

Apparent improvements in thresholds at retest are more likely at low stimulus frequencies and in cases where baseline thresholds are poorer. The effect of baseline threshold could be a combination of audiometer and/or noise floor artifacts (e.g., a baseline threshold of −5 dB HTL has a limited range of potential improvement and a practically unlimited range of potential decrement) and changes that are not artifacts of the measurement process.

This study invited volunteers across a wide range of age and hearing sensitivity, but we had small numbers of participants with poor baseline thresholds in the low- and mid-frequencies because the study sample was drawn from the general population where such impairments are comparatively uncommon (Ciletti & Flamme, 2008). In addition, the rate of low- and mid-frequency hearing impairment should be lower due to the exclusion of participants with signs and symptoms of active middle ear disorders. This sampling design was useful for the purpose of examining the variability of thresholds as they are distributed in the general population, but a separate study of people with no thresholds better than about 20 dB HTL at baseline would be required to minimize the role of measurement artifact on observed threshold differences.

To our knowledge, the role of WBR measures on pure tone threshold differences has not been shown before in a sample without signs or symptoms of middle ear dysfunction. The association with WBR results at baseline was small, but controls for the average differences in middle ear efficiency as a function of frequency via the inclusion of that factor in the analytic model. While confirmation in a separate study is desired, it appears that normally-functioning ears reflecting more energy at a given frequency are more likely to show apparent decrements in hearing sensitivity on retest.

The effect of changes in WBR on test-retest differences is both larger and conceptually simpler. Reductions in the efficiency of middle-ear energy transfer can be expected to reduce cochlear stimulation and therefore lead to an apparent worsening of hearing sensitivity. The presence of this effect is remarkable from two perspectives. First, the effect was found among participants with no obvious signs or symptoms of middle ear dysfunction, and second, the effect was better represented by ambient-pressure WBR than the conventional tympanometric measures obtained in this study (i.e., compliance and peak pressure). The tympanometers used in this study were designed for screening purposes, so it remains possible that a system designed for differential diagnosis could return better data than the screening units. However, analyses of the WBR data obtained in this study revealed that there is scant correlation between mid- and high-frequency energy reflectance and reflectance observed at 226 Hz (Flamme et al., 2013). Therefore a conventional 226 Hz tympanogram should not be expected to yield much information generalizable to the high frequencies in ears without evidence of dysfunction.

The effect of recent noise exposure was small but in the expected direction. Among the noise exposure variables included in this study, the strongest predictor of test-retest difference was the average sound level during the eight hours preceding each measurement visit, which is consistent with prior research (Nixon et al., 1977). The interquartile range of 8-hour average noise exposures spanned 10 dB (64 to 74 dB L_{Aeq8}), so the marginal difference in expected test-retest difference across the middle 50 % of the sample was less than 0.2 dB after the coefficient in Table 3 is applied.

It seems natural to question whether the 12-hour quiet period recommended prior to occupational testing (NIOSH, 1998) is justified given the present data, but it is important to recognize that this study sample was drawn from the community, in which fewer than 20 % of men and 10 % of women had 8-hour equivalent daily exposures exceeding the NIOSH Recommended Exposure Limit (Flamme, Stephenson et al., 2012). The effects of recent noise exposure, with and without accounting for hearing protector effects, should be studied among participants in occupational hearing conservation programs. All participants in a noise-based occupational hearing conservation program would be expected to have exposures exceeding the NIOSH exposure limit, so any observed relationship between noise exposure and test-retest threshold differences lends support to a mandatory quiet period to minimize the effects of temporary threshold shift.

The strong association between age and test-retest difference is difficult to interpret in a statistical model that includes baseline thresholds. The trend toward apparent worsening of thresholds at retest began among participants in their 30s and increased with age. Age often acts as a proxy for other factors, and a proxy relationship cannot be ruled out except for the variables included in the univariable analyses (Appendix) that either showed no relationship with test differences or were subsequently dropped from the multivariable model in favor of variables more strongly related to the outcome. Additional study is needed to determine the nature of the age differences observed here.

A relationship between the characteristics of the tester and test-retest differences was not expected in a study that utilized scripted and recorded instructions, automated threshold test

procedures, and a rigid protocol for earphone placement. This effect seems most likely due to a subject-experimenter artifact (Rosenthal & Rosnow, 1991) (p 110–134), despite the employment of multiple strategies (e.g., restricted and standardized communication with participants, low-keyed and nonthreatening procedures) to reduce artifacts. Future studies on the magnitude and direction of tester influences are warranted.

Biases and limitations

This large-scale study was designed to inform decisions about the utility of mandatory inclusion of thresholds at 8 kHz in occupational hearing conservation programs. As such, the study was able to provide precise estimates of test-retest differences and high statistical power to detect correlates of these differences. However, the difference distributions obtained in this study cannot be generalized beyond the transducer (TDH-39P), stimuli (pure tones), and automated threshold algorithm (modified Hughson-Westlake, as implemented in the Tremetrics HT Wizard) employed in the study.

The test-retest variability observed in this study might not generalize to other transducer styles. For example, prior work with insert and circumaural earphones suggested that these earphone styles might produce less test-retest variability (Frank, 2001; Schmuziger et al., 2004) than observed with the TDH-39P supra-aural earphones used in this study. Furthermore, changes in the temporal or spectral characteristics of the stimulus could reduce the likelihood of interactions between the transducer and the ear, or that different test algorithms could be employed to yield more stable thresholds.

The participants in this study knew they were taking part in research and were given nominal reimbursement for their time and inconvenience, and these factors could change the participants' level of motivation (Dobie, 1983) or other aspects of participant behavior such as responsiveness to the demand characteristics of the study protocol (Rosenthal & Rosnow, 1991). Although we strove to minimize participant distractions and conducted a swift protocol wherein both audiometric tests were conducted in less than 15 minutes, these factors cannot be ruled out entirely and should be examined in future studies. However, one might expect that workers and research participants would both experience a decline in interest and motivation as the measurement visit progressed, which would lead to increased test-retest differences for thresholds obtained later in the measurement visit; this effect would engender a significant effect of tests within measurement visits. No such effect was observed in this study. However, if an overall motivation effect were present in field settings, that effect would likely deliver an additional source of overall variance that would affect all frequencies equally.

The daily calibration procedures employed in this study (e.g., twice-daily measurement of SPL developed in an ear simulator) would be atypical in occupational settings; however the use of bio-acoustic simulators is common in occupational audiometry settings. Minimal calibration drift over time was observed in this study, suggesting that consistent use of a modern bio-acoustic simulator would monitor for hardware failures adequately.

The periodic comprehensive calibration using an IEC-60318-1 ear simulator was also atypical. The NBS-9A cylindrical coupler is more commonly used with the TDH-39P

earphone. However, the NBS-9A was never intended to simulate the impedance properties of the human ear, and it was first used during the 1930s after completion of the National Health Study (Corliss & Snyder, 1950). The diameter of the NBS-9A cylinder was designed to accommodate a 1-inch measurement microphone. A large diameter microphone is problematic at high frequencies, where the dimensions are large relative to stimulus wavelength and uniform air particle motion is not guaranteed. Furthermore, there is a report of a physical interaction between the TDH-39P earphone and the NBS-9A coupler that leads to calibration errors for some earphone samples (Lutman & Qasem, 1998; Flamme & Tatro, unpublished data, 2010). The physical design of the IEC-60318-1 (IEC, 2009) ear simulator accommodates higher frequencies and its use is currently advocated for both supra-aural and circumaural earphones (ANSI, 2010).

Conclusion

The test-retest variability of pure tone thresholds at 6 and 8 kHz is poorer than at other frequencies, and 8 kHz is slightly worse than 6 kHz. However, the difference has no impact on the 50 % or 90 % critical difference and the joint variability can be expected to only rarely result in a spurious 6 kHz notch of 10 dB or greater. The consequence of missing the majority of audiometric notches, which are centered on 6 kHz and require a threshold measurement at 8 kHz for detection as a notch, is far more problematic than the small increase in false positive responses that would result from assessment of thresholds through 8 kHz. In addition to stimulus frequency, other factors leading to significant predictions of signed changes in threshold were baseline threshold, baseline WBR, change in WBR, noise exposure, age, and the level of tester education.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank CAPT William J. Murphy, Ph.D., Peter B. Shaw, Ph.D. (CDC/NIOSH/Division of Applied Research and Technology), Elliott Berger, M.S. (3M Personal Safety Division, Indianapolis, IN), Jack Forman (Tremetrics, Eden Prairie, MN), Kim Schairer, Ph.D. (VAMC-Mountain Home, TN), John Brown, Chas Pudrith, Lydia Baldwin, Emma Trabue, Joyce Gard, and Hannah Borton (WMU) for their assistance with this project.

This study was supported by CDC/NIOSH Contract 211-2009-31218.

Abbreviations

ANSI	American National Standards Institute
ASA	Acoustical Society of America
CDC	US Centers for Disease Control and Prevention
CI	Confidence interval
dB	decibel
dBA	A-weighted decibel

HTL	Hearing Threshold Level
Hz	Hertz
IEC	International Electrotechnical Commission
kHz	kilohertz
L_{Aeq}	A-weighted equivalent continuous sound pressure level
L_{Aeq8}	A-weighted 8 hour equivalent continuous sound level
NBS	US National Bureau of Standards
NHANES	US National Health and Nutrition Examination Survey
NIHL	noise-induced hearing loss
NIOSH	National Institute for Occupational Safety and Health
OHC	Occupational hearing conservationist
OSHA	Occupational Safety and Health Administration
SEM	Structural equation model
SPL	sound pressure level
TDH	Telephonics Dynamic Headphone
WBR	Wideband reflectance
WBT	Wideband tympanometry
WMU	Western Michigan Universiry

References

ANSI S1.4-1983. American National Specification for Sound Level Meters. Melville, New York: Acoustical Society of America; R2006.

ANSI S3.44-1996. Determination of occupational noise exposure and estimation of noise-induced hearing impairment. Melville, New York: Acoustical Society of America; R2006.

ANSI S3.1-1999. Maximum Permissible Ambient Noise Levels for Audiometric Test Rooms. New York, NY: American National Standards Institute, Inc; R2008.

ANSI S3.6-2004. Specification for audiometers. Melville, New York: Acoustical Society of America;

ANSI S3.6-2010. Specification for Audiometers. Melville, New York: Acoustical Society of America;

Atherley GR, Dingwall-Fordyce I. The Reliability of Repeated Auditory Threshold Determination. Br J Ind Med. 1963; 20:231–235. [PubMed: 14046161]

Atherley GR, Johnston N. Audiometry - the ultimate test of success? Ann Occup Hyg. 1981; 27:427–447. [PubMed: 6660687]

Carhart R, Hayes C. Clinical Reliability of Bone Conduction Audiometry. Laryngoscope. 1949; 59:1084–1101. [PubMed: 15406725]

Carhart R, Jerger JF. Preferred Method for Clinical Determination of Pure-Tone Thresholds. J Speech Hear Disord. 1959; 24:330–345.

Ciletti L, Flamme GA. Prevalence of Hearing Impairment by Gender and Audiometric Configuration: Results from the National Health and Nutrition Examination Survey (1999–2004) and the Keokuk County Rural Health Study (1994–1998). J Am Acad Audiol. 2008; 19:672–685. [PubMed: 19418707]

Coles RR, Lutman ME, Buffin JT. Guidelines on the diagnosis of noise-induced hearing loss for medicolegal purposes. Clin Otolaryngol. 2000; 25:264–273. [PubMed: 10971532]

Corliss ELR, Snyder WF. Calibration of audiometers. J Acoust Soc Am. 1950; 22:837–842.

Dobie RA. Reliability and validity of industrial audiometry: implications for hearing conservation program design. Laryngoscope. 1983; 93:906–927. [PubMed: 6865627]

Dobie RA. Estimating noise-induced permanent threshold shift from audiometric shape: the ISO-1999 model. Ear Hear. 2005; 26:630–635. [PubMed: 16377998]

Flamme, GA.; Deiters, K.; Tatro, A.; Geda, K.; McGregor, K. Reliable differences in wideband otoreflectance patterns among adults. American Auditory Society Annual Meeting; Scottsdale, AZ. 2013.

Flamme GA, Stephenson MR, Deiters K, Tatro A, VanGessel D. Typical noise exposure in daily life. Int J Audiol. 2012; 51:S3–S11. [PubMed: 22264061]

Frank T. High-frequency (8 to 16 kHz) reference thresholds and intrasubject threshold variability relative to ototoxicity criteria using a Sennheiser HDA 200 earphone. Ear Hear. 2001; 22:161–168. [PubMed: 11324845]

Gravendeel DW, Plomp R. The relation between permanent and temporary noise dips. Arch Otolaryngol. 1959; 69:714–719.

Guild SR. A method of classifying audiograms. Laryngoscope. 1932; 42:821–836.

Hetu R. Critical analysis of the effectiveness of secondary prevention of occupational hearing loss. J Occup Med. 1979; 10:36–44.

High WS, Gallo RP. Audiometric reliability in an industrial hearing conservation program. J Aud Res. 1963; 3:15–34.

Hosmer, DW.; Lemeshow, S. Applied Logistic Regression. Hoboken, NJ: John Wiley & Sons; 2000.

Huber, PJ. The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the Fifth Berkely Symposium on Mathematical Statistics and Probability; University of California Press; Berkely, CA. 1967. p. 221-233.

IEC. IEC-60318-1 Electroacoustics – Simulators of the human head and ear – Part 1: Ear simulator for the measurement of supra-aural and circumaural earphones. Geneva, Switzerland: IEC; 2009.

Lutman, ME.; Qasem, HYN. A source of notches at 6 kHz. In: Prasher, D.; Luxon, L., editors. Advances on Noise Research: Biological Effects of Noise. 1998. p. 170-176.

NIOSH. US Department of Health and Human Services. Cincinnati, OH: National Institute for Occupational Safety and Health; 1998. Occupational Noise Exposure: Revised Criteria 1998. DHHS Publication 98-126.

Nixon CW, Johnson DL, Stephenson MR. Asymptotic behavior of temporary threshold shift and recovery from 24- and 48-hour exposures. Aviat Space Envir Med. 1977; 48:311–315.

OSHA. Occupational Noise Exposure ; Hearing Conservation Amendment, part two of two. US Department of Labor, Occupational Safety and Health Administration. 1981:34. 29 CFR Part 1910, 46 Federal Register 4078.

OSHA. Occupational noise exposure : Hearing Conservation Amendment; Final rule. US Department of Labor, Occupational Safety and Health Administration. 1983 29 CFR 1910.95, 48 Federal Register 9776-9785.

OSHA. [August 19, 2013] Recording Criteria for Cases Involving Occupational Hearing Loss. US Department of Labor, Occupational Safety and Health Administration. 2003. Retrieved from: http://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=STANDARDS&p_id=9641

Rabinowitz PM, Galusha D, Slade MD, Dixon-Ernst C, Sircar KD, et al. Audiogram notches in noise-exposed workers. Ear Hear. 2006; 27:742–750.

Rosenthal, R.; Rosnow, RL. Essentials of Behavioral Research: Methods and Data Analyses. New York, NY: McGraw-Hill; 1991.

Schmuziger N, Probst R, Smurzynski J. Test-retest reliability of pure-tone thresholds from 0.5 to 16 kHz using Sennheiser HDA 200 and Etymotic Research ER-2 earphones. Ear Hear. 2004; 25:127–132. [PubMed: 15064657]

Stephenson MR, Themann CL, Murphy WJ. Thoughts on the noise “notch” and the importance of testing 8 kHz. CAOHC Update. 2004; 16(3):1, 6.

Tukey, JW. Exploratory Data Analysis. Reading, MA: Addison-Wesley; 1977.

US Department of Labor. [August 19, 2013] Bureau of Labor Statistics Standard Occupational Classification. US Department of Labor, Bureau of Labor Statistics. 2010. Retrieved from: www.bls.gov/soc/

Veneman CE, Gordon-Salant S, Matthews LJ, Dubno JR. Age and Measurement Time-of-Day Effects on Speech Recognition in Noise. Ear Hear. 2013; 34:288–299. [PubMed: 23187606]

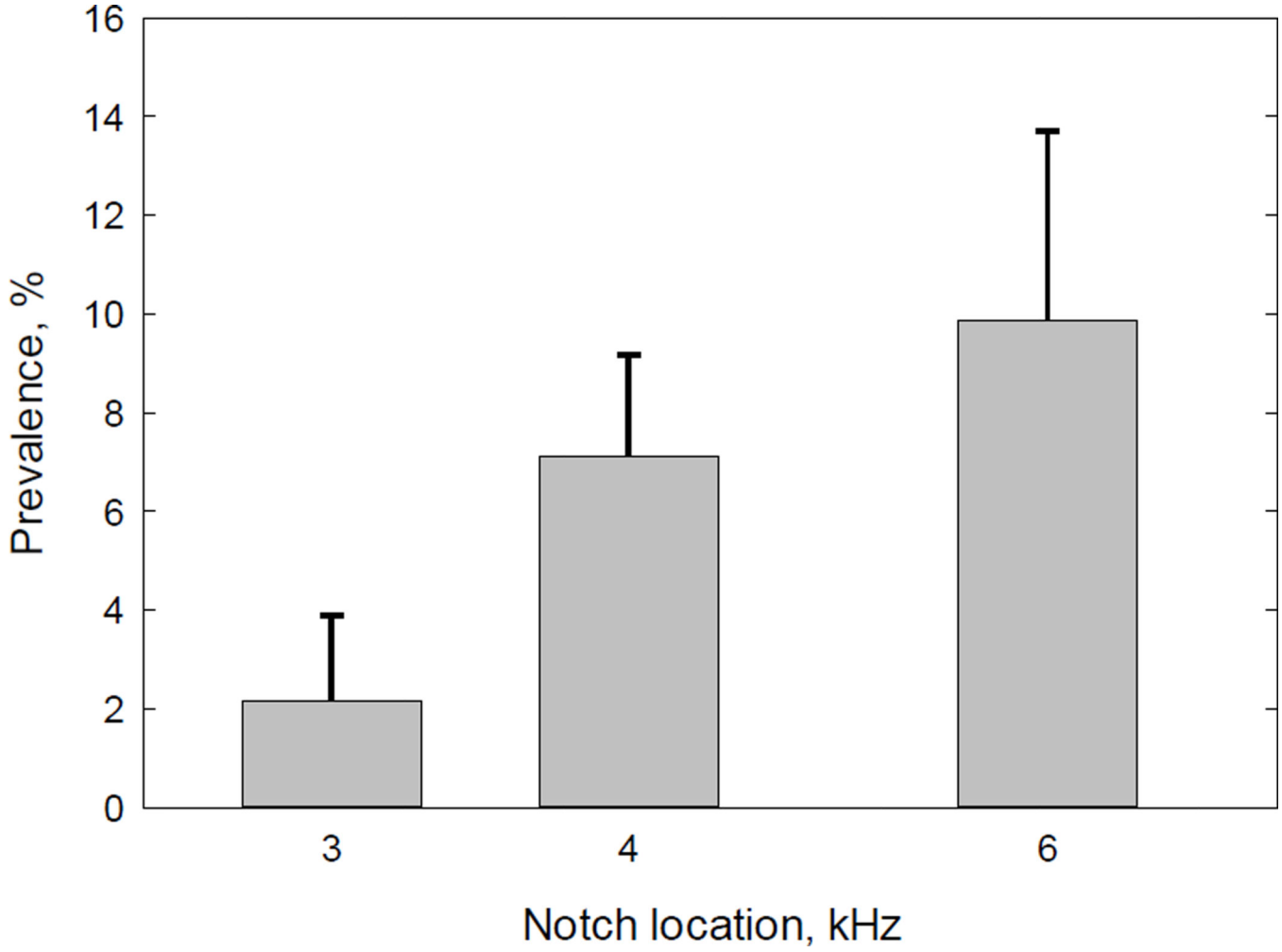

Figure 1.

Prevalence of notch locations, in kHz, among people exposed to occupational noise more than 3 months. Unpublished analysis of NHANES 1999–2004 data, ages 20–69. Notch definition: 15 dB depth relative to the average of 0.5, 1, and 2 kHz and 15 dB improvement at 8 kHz in relation to the worst threshold. Error bars represent the 95 % confidence interval for the percentage, accounting for the NHANES sample stratification, clustering, and weighting.

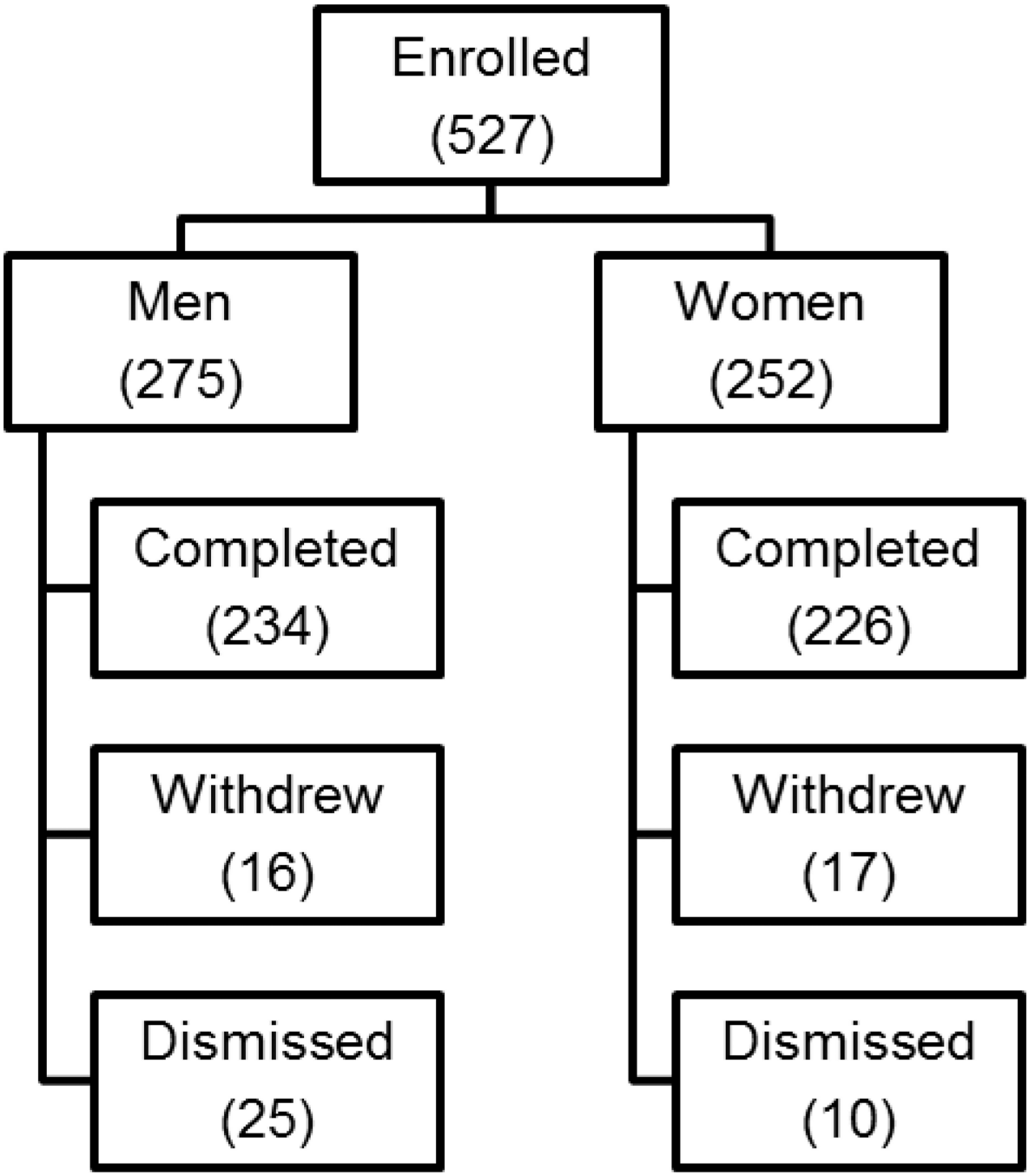

Figure 2.
Participant flow

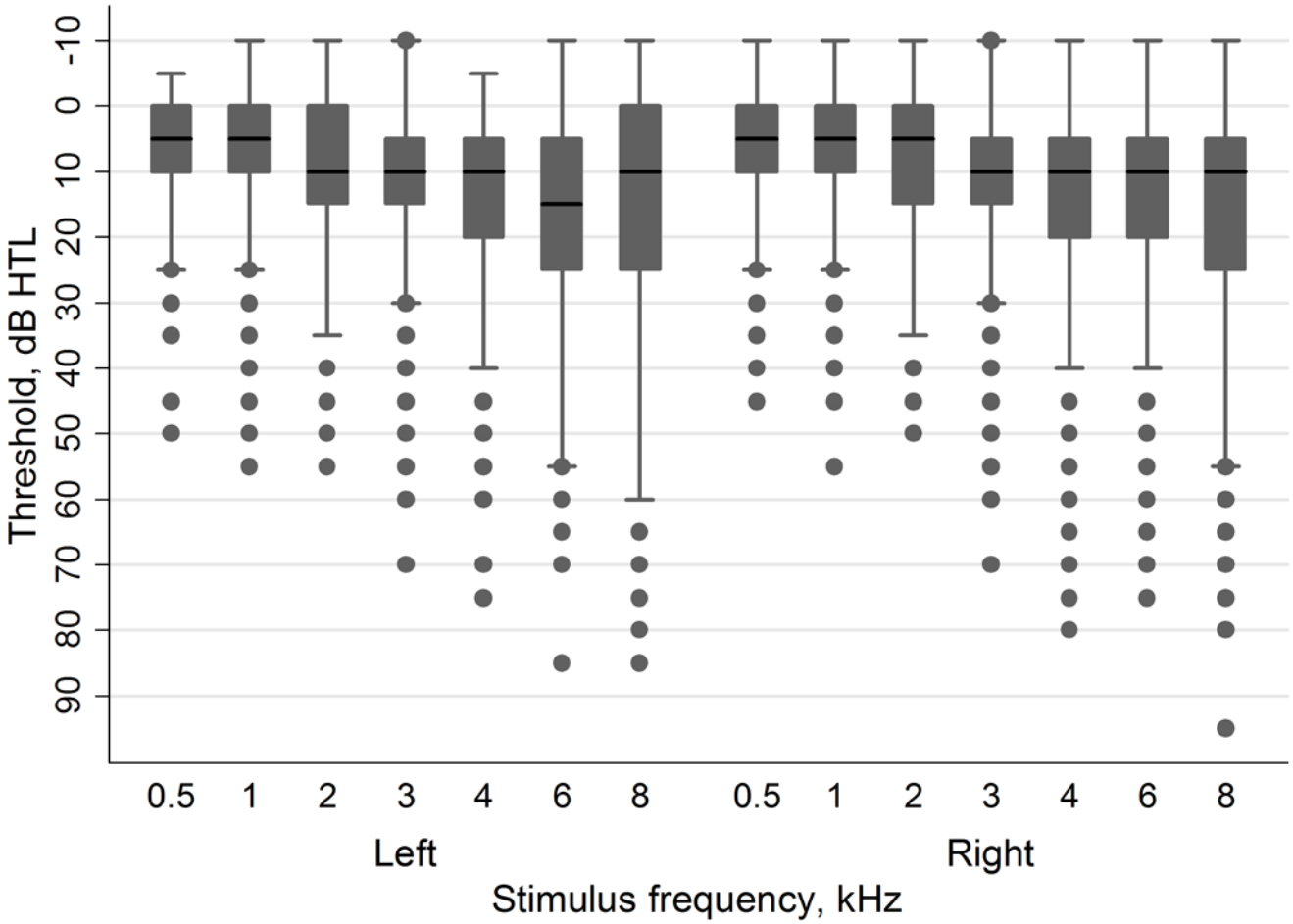

Figure 3.

Pure tone thresholds at baseline (i.e., the first test on the first visit), by ear. Bold black lines in boxes represent medians; shaded regions represent the interquartile range. Error bars represent the upper and lower adjacent values (Tukey, 1977); filled circles represent values falling outside the upper and lower adjacent values.

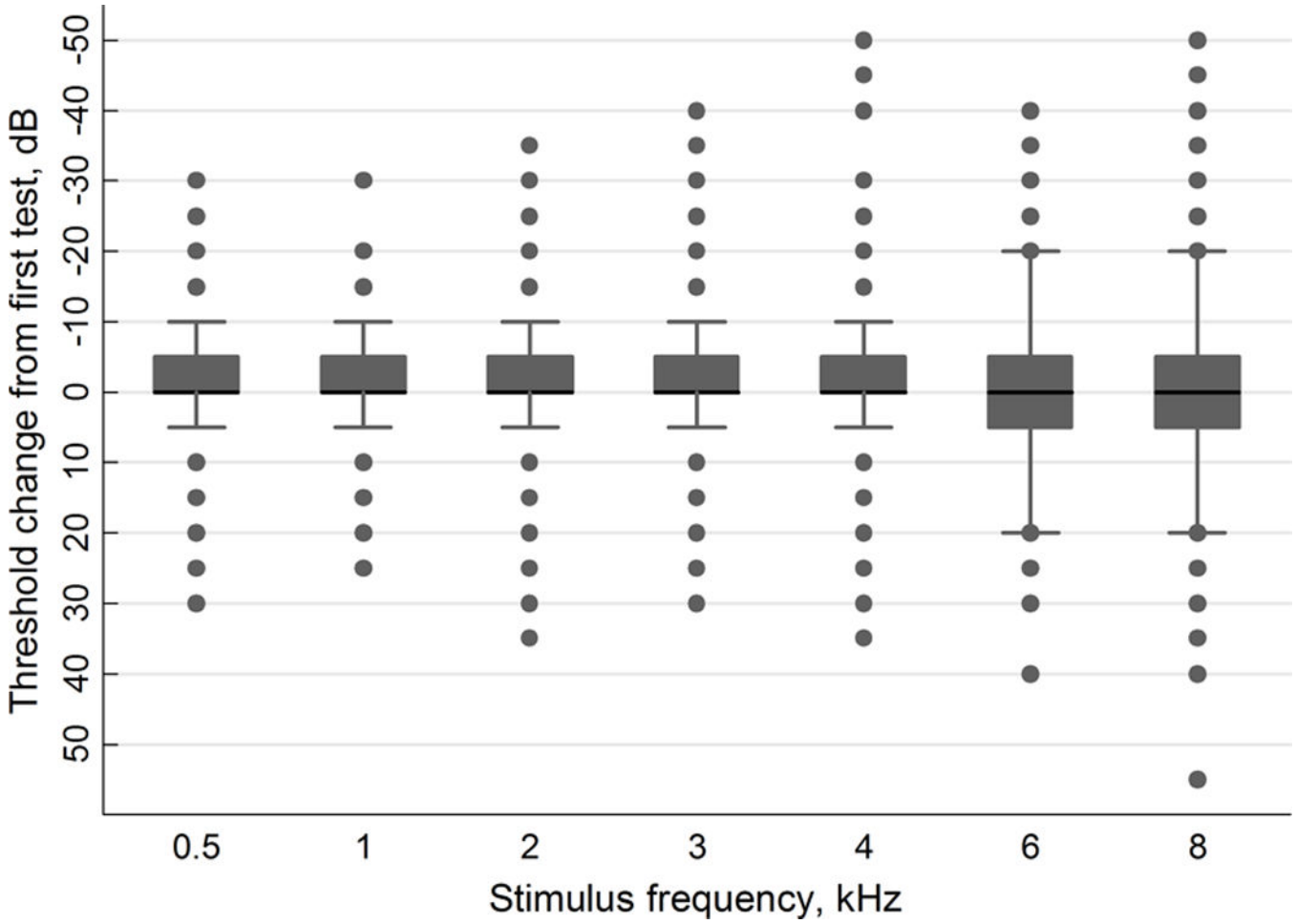

Figure 4.

Boxplot representing test-retest difference distributions across stimulus frequencies. Box details are identical to Figure 3.

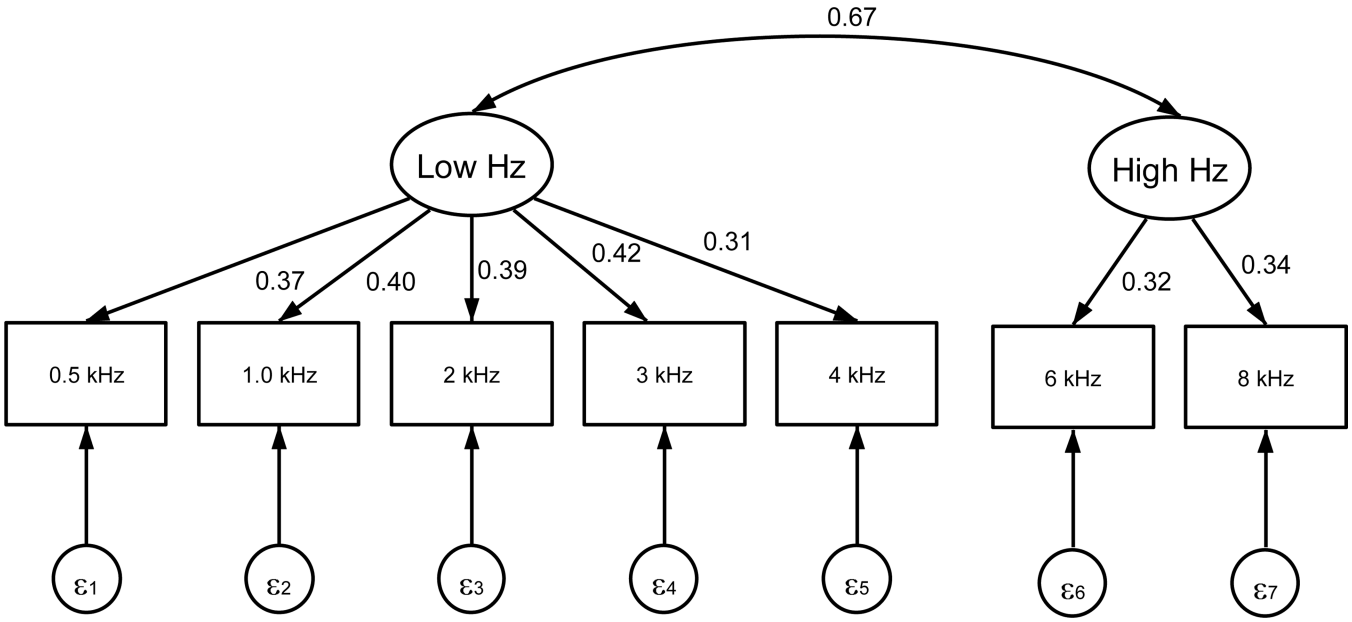

Figure 5.

Structural equation model representing the correlation structure for test-retest differences. Squares represent observed variables. Ovals represent latent variables comprised by the associated observed variables. Coefficients associated with straight arrows represent the correlations between latent and observed variables. The curved arrow represents the correlation between latent factors.

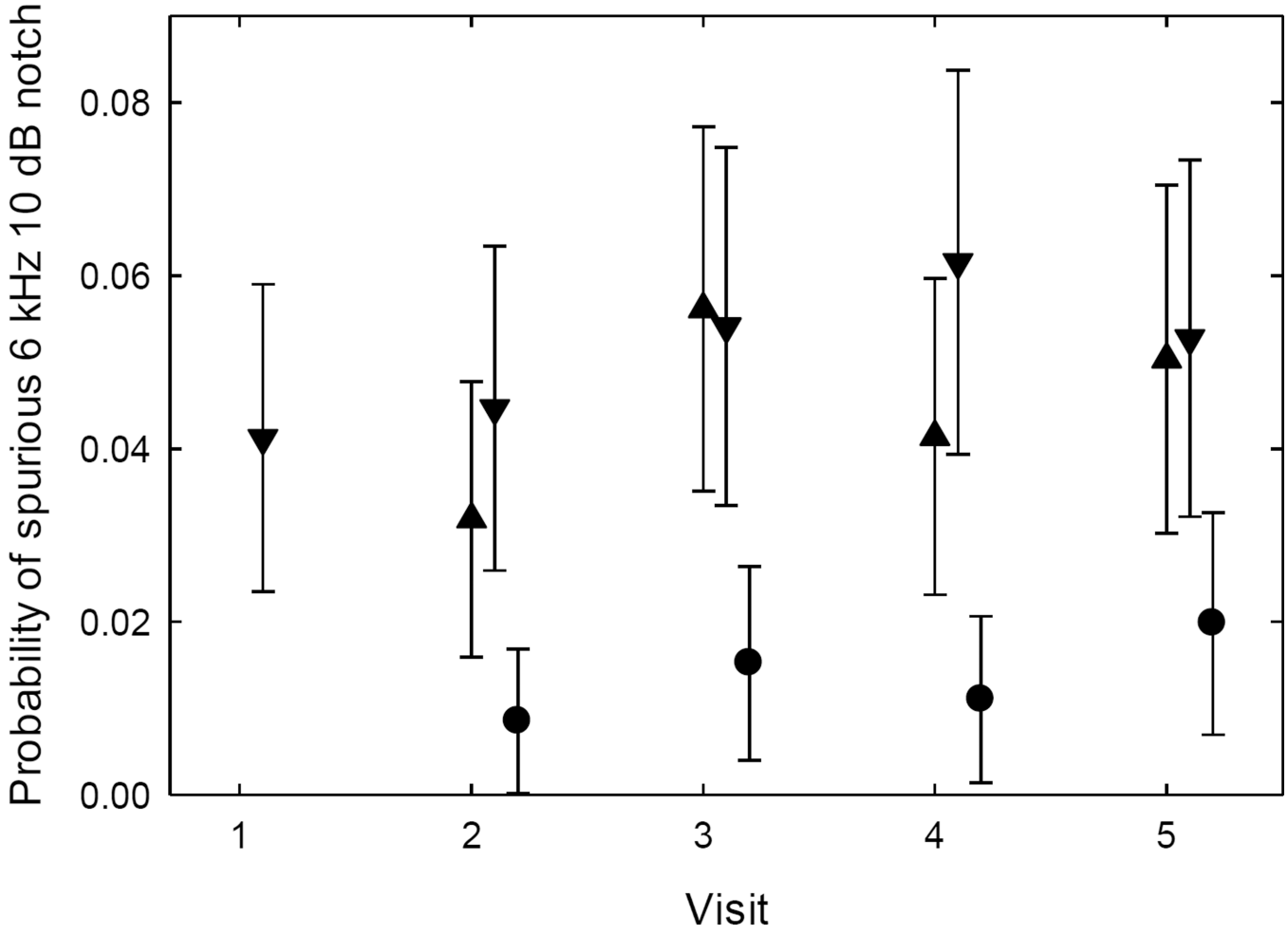

Figure 6.

Probabilities of spurious 10 dB notches in left ears at 6 kHz. Upward and downward triangles represent the first and second test within a measurement visit, respectively. Circles represent consecutive tests within a measurement visit. Error bars represent the 95 % confidence intervals for each proportion. An erroneous notch was defined as a deviations of +5 dB (worse) or more at 6 kHz and a −5 dB deviation or more at one or more of 2, 3, or 4 kHz and a −5 dB deviation or more at 8 kHz, which would yield an apparent 10 dB notch at 6 kHz due to measurement error.

Table 1

Descriptive statistics and short-term critical differences for threshold variables.

			Critical Difference boundaries		
Frequency, kHz	**Mean**	**SD**	**50 %**	**80 %**	**90 %**
0.5	−1.1	4.6	[−5, 0]	[−5, 5]	[−10, 5]
1	−0.7	4.0	[−5, 0]	[−5, 5]	[−5, 5]
2	−0.8	4.4	[−5, 0]	[−5, 5]	[−5, 5]
3	−0.8	4.6	[−5, 0]	[−5, 5]	[−5, 5]
4	−0.5	5.3	[−5, 0]	[−5, 5]	[−10, 5]
6	−0.5	6.4	[−5, 5]	[−10, 5]	[−10, 10]
8	−0.2	7.0	[−5, 5]	[−10, 10]	[−10, 10]

Table 2

Correlations between test-retest differences across stimulus frequency.

Frequency, kHz	0.5	1	2	3	4	6
1	0.33					
2	0.16	0.23				
3	0.20	0.23	0.26			
4	0.10	0.14	0.22	0.25		
6	0.08	0.07	0.10	0.15	−0.01*	
8	0.12	0.11	0.11	0.14	0.05	0.25

* Not significant ($p > 0.05$). All other correlations were significant at the 0.05 level.

Table 3

Multilevel regression model predicting signed test-retest differences.

	Coefficient	Robust Std. Err.	*z*	*p*	95 % CI	
0.5 kHz stimulus, binary	−2.446	0.211	−11.60	< 0.0005	−2.856	−2.031
1.0 kHz stimulus, binary	−1.900	0.203	−9.34	< 0.0005	−2.297	−1.500
2.0 kHz stimulus, binary	−1.542	0.207	−7.46	< 0.0005	−1.945	−1.137
3.0 kHz stimulus, binary	−1.258	0.244	−5.95	< 0.0005	−1.673	−0.844
4.0 kHz stimulus, binary	−0.618	0.238	−2.60	< 0.0005	−1.086	−0.151
6.0 kHz stimulus, binary	−0.872	0.213	−4.10	< 0.0005	−1.287	−0.455
Threshold at baseline, dB	−0.133	0.011	−12.21	< 0.0005	−0.154	−0.112
WBR (dB) at baseline	0.074	0.029	2.58	0.010	0.018	0.129
WBR (dB) change from baseline	0.136	0.033	4.14	< 0.0005	0.072	0.202
Average noise exposure during last 8 hours, dB	0.017	0.007	2.48	0.013	0.004	0.030
Age 30–39, binary	1.078	0.246	4.39	< 0.0005	0.597	1.560
Age 40–49, binary	1.548	0.257	6.01	< 0.0005	1.043	2.052
Age 50–59, binary	2.491	0.258	9.65	< 0.0005	1.98	2.997
Age 60–69, binary	3.819	0.411	9.30	< 0.0005	3.014	4.623
Graduate–level audiology education for tester, binary	0.319	0.107	2.99	0.003	0.110	0.528
Constant	−0.509	0.513	−0.99	0.321	−1.515	0.496